Robotics and Autonomous Systems 61 (2013) 1460–1475

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

Learning spatially semantic representations for cognitive robot navigation

Ioannis Kostavelis *, Antonios Gasteratos

Laboratory of Robotics and Automation, Production and Management Engineering Department, Democritus University of Thrace, 12, Vas. Sofias Str, GR-671 00, Xanthi, Greece

HIGHLIGHTS

- Two level navigation.
- Cognitive navigation.
- Spatial semantics.

ARTICLE INFO

Article history:
Received 6 December 2012
Received in revised form
7 July 2013
Accepted 15 July 2013
Available online 22 July 2013

Keywords:
Semantic mapping
Spatial visual memories
Place classification
SLAM
Neural Gas
Bag-of-features
Topological graph

ABSTRACT

Contemporary mobile robots should exhibit enhanced capacities, which allow them self-localization and semantic interpretation as they move into an unexplored environment. The coexistence of accurate SLAM and place recognition can provide a descriptive and adaptable navigation model. In this paper such a two-layer navigation scheme is introduced suitable for indoor environments. The *low layer* comprises a 3D SLAM system based solely on an RGB-D sensor, whilst the *high* one employs a novel content-based representation algorithm, suitable for spatial abstraction. In course of robot's locomotion, salient visual features are detected and they shape a *bag-of-features* problem, quantized by a Neural Gas to code the spatial information for each scene. The learning procedure is performed by an SVM classifier able to accurately recognize multiple dissimilar places. The two layers mutually interact with a semantically annotated topological graph augmenting the cognition attributes of the integrated system. The proposed framework is assessed on several datasets, exhibiting remarkable accuracy. Moreover, the appearance based algorithm produces semantic inferences suitable for labeling unexplored environments.

1. Introduction

Along the last decades persistent research endeavors in the areas of robotics and artificial intelligence revealed the great challenge of cognitive navigation, an area that combines robot mobility with high level perception of the environment and it can be distinguished into two discrete primitives [1]. The first one involves the *numerical navigation*, that is the capacity of the robots to localize themselves and generate a metric map of the explored environment. The second primitive involves the *semantic interpretation*, which encompasses the competence of the robots to apprehend their environment. The latter employs appearance based features of the explored space and along with machine learning techniques allows the artificial agents to perform visual place classification. Consequently, such an augmented navigation framework entails an enhanced human–robot interaction and communication. A more effective insight about the significance of the cognitive navigation is possible by assuming the paradigm of an "office robot", in charge of fetching objects. The robot will efficiently accomplish this missions only if it can manage to accurately estimate its current position and to understand spatially semantic representations in consecutive time instances. Thereupon, a fundamental communication axis between the human and the robot passes through the ability of the second to perceive its proper environment and, wherefore, to accurately recall learned spatial memories. The latter is an important challenge, inasmuch as vision based place recognition and categorization systems demand great abstraction of the spatial representation due to the vast amount of information they take in. Following the human ingenuity, a sought competence for a cognitive robot should be to make robust semantic inferences based on context interpretation mechanisms, even when visiting places for the first time.

Notwithstanding the plethora of laborious research conducted in the specific field [2], the majority of the proposed works tackle only a fraction of cognitive navigation [3,4]. The paper at hand aims to outline the basic objectives of a complete cognitive navigation

* Corresponding author. Tel.: +30 2541079330; fax: +30 2541079331.
E-mail addresses: gkostave@pme.duth.gr, kostavelis@gmail.com (I. Kostavelis), agaster@pme.duth.gr (A. Gasteratos).

0921-8890/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.robot.2013.07.008

framework, while it introduces efficient novel solutions to each of them. The entire problem has been decomposed in discrete tasks by organizing the key-parts of an efficient cognitive navigation framework as follows:

- *Objective* 1: The *low layer* navigation. It comprises all the numerical and geometrical attributes that a robot should retain for its localization. The accurate location estimations and the formation of a consistent map of the explored area along the course of the robot in an unknown environment constitute the cornerstones of a successful navigation scheme.
- *Objective* 2: The *high layer* navigation. It involves all the cognitive attributes that a robot should possess to effectively draw semantic inferences about its current location. This objective can be further broke down into several sub-objectives.
  1. *Spatial abstraction*. The system should be able to store in an abstract and representative manner any learned spatial information, which is gathered in course of the exposure of the robot into newly visited places. That is, the system should retain its recall capabilities no matter how many different places has to learn, e.g. "office", "corridor", "bathroom", "kitchen", "stairs", etc.
  2. *Place recognition*. The system should be able to accurately recognize the learned place instances as it moves from one location to another. The latter requires the utilization of robust machine learning techniques competent to deal with any dynamic change of the explored environments.
  3. *Place categorization*. The system should be proficient in categorization and not only in recognition of different places. That is, the robot should be capable of classifying and producing labels for places about which no prior knowledge is available. In other words, the system should be able to generalize the knowledge gained by exploring a specific spot, so as to infer about the semantic content of any other similar place.
- *Objective* 3: A connectivity framework interfacing the high and low operations. It should be responsible for the coordination, of the aforementioned layers enabling the information exchange among them by combining both geometrical and semantic inferences about the robot's surroundings.

Towards the fulfillment of these objectives, the proposed work outlines, implements and presents a compound *two-layer* navigation framework, while it proposes a hybrid intermediate layer responsible for their coordination. More specifically, at the *low layer* of the navigation scheme a Simultaneously Localization and Mapping (SLAM) algorithm was developed suitable only for indoor exploration, due to the fact that it is solely based on a RGB-D sensor, to wit the Microsoft Kinect. The proposed SLAM algorithm exploits the depth information acquired directly from this sensor and by utilizing robust features it calculates the incremental motion estimation between the successive time instances. The solution to the motion estimation computational problem is provided by the "Procrustes" analysis. At the same time a dense point cloud of the scene is calculated. In an ideal scenario, the robot's motion estimation and the point cloud at each successive frame should be sufficient to form the map of the explored area. However, in the real case, an intermediate refinement step is required to precisely merge the point clouds. Therefore, in each step a plane detection RANSAC algorithm is applied on the point cloud delivered by the RGB-D at the respective time instance, which distinguishes the most prominent surfaces in the scene. The points that belong to the same detected plane among two consecutive frames share a great amount of spatial information and, therefore, they are selected to be fed into an Iterative Closest Point (ICP) algorithm that refines the point cloud merging procedure. The output of this algorithm is a coherent map of the explored area. The *high layer* navigation scheme holds the major novelty in the proposed work. It empowers an effective spatial abstraction of the explored space followed by the memorization of the distinct places such as "office", "corridor", etc. More specifically, the system has to learn a very abstract representation of the visited areas, wherefore for each time instance the scene features are detected and stored. Consequently, we end up with an excessive number of rotation, scale and illumination invariant features that describe all the place instances need to be learned. This abundant of information is treated as a *bag-of-features* problem [5]. Specifically, all the extracted features are fed into a Neural Gas, which is a space quantization algorithm that ends up with a sparse set of representation vectors abstracting the initial input space. In the next step, the detected features in each view are spread over the calculated quantization vectors to form appearance based histograms, which comprise a content based representation of the initial space and are utilized in the off-line training procedure of a Support Vector Machine (SVM) classifier. Both the two layers are seamlessly connected by means of a semantically annotated topological graph. The latter is formed exploiting the odometry output of the *low layer*, while it also takes advantage of semantic attributes from the *high layer* navigation scheme. This intermediate step retain visual memories as nodes in the topological graph, which are further utilized to differentiate among places of the same type. A further contribution of the proposed work is the introduction of a dataset suitable for indoor localization and semantic mapping, which has been captured by our MAGGIE (Mobile Autonomous riGGed Indoors Exploratory) robot (Fig. 1). The suggested cognitive navigation integrated framework has been tested by means of this dataset and exhibited remarkable accuracy. Additionally, the place classification (i.e. recognition and categorization) capabilities of the *high layer* mapping algorithm are further examined on another public available dataset [6].

The rest of the paper is organized as follows: in Section 2 a review of the related literature is provided. In Section 3 the *two-layer* cognitive navigation framework including the intermediate semantic topological graph is presented, while in Section 4 the new Cognitive Navigation dataset is introduced. Section 5 describes the experimental evaluation proving the significance of the proposed method. In Section 6 conclusions about the presented work are drawn.

2. Related work

Cognitive navigation methods include a significant variety of implementations each of which might make use of several sensory input [7]. Due to the fact that many navigation algorithms with cognition characteristics have been proposed in the literature, an attempt to provide an appropriate categorization of all the trends and the particularities reveal certain difficulties. A common separating line among the existing techniques is the utilization of different sensory input for the construction of the 3D map. In particular, methods designed to learn 3D maps of the environment employ laser scanners or Time-of-Flight (ToF) cameras to provide dense point clouds of the environment [8,9]. Other common sensors are the RGB-D and stereo cameras. The SLAM approaches that utilize RGB-D images are intrinsically different from stereo systems as their input is directly a dense point cloud instead of a pair of color images [10,11]. Such an information rich sensory output can be used not only to reconstruct a 3D map, but to provide 3D object segmentation and recognition [12], which is an essential step toward the building of an integrated framework to perceive the environment from an artificial agent. Typically, the visual SLAM systems are also cited in the literature as *structure from motion* ones [13] that extract and track sparse keypoints from the camera images. The most common methods to obtain reliable keypoints are the SIFT [14] and SURF [15] ones, while in the literature one can refer to more recent approaches such as the CenSurE [16], a non-linear Difference of Gaussian SIFT-based method [17], or concurrent versions such as the SIFTGPU [18]. In a more sophisticated

Fig. 1. The MAGGIE (Mobile Autonomous riGGed Indoors Exploratory) robot.

approach, the robot localization is aided by detecting objects in a scene [19]. The RGB-D sensors that are based on structured light, directly provide dense depth maps and color images. An example of this method is described in [11], where a SLAM system utilizes bundle adjustment to align the dense point clouds acquired directly by a Kinect sensor without any use of the RGB images. Moreover, Henry et al. in [20] proposed a sparse keypoint matching among successive color images as an initialization to an ICP algorithm and proved that the often utilization of the computationally expensive ICP step was not necessary. Regarding linearized EKF-based approach, in [21] a First Estimates Jacobian EKF is proposed to improve the estimator's consistency during SLAM.

An alternative way to distinguish among cognitive navigation paradigms is the method utilized to manage the spatial knowledge of the system. An efficient way to organize the information about a place or an object is by treating the issue as a *bag-of-words* (also known in computer vision as a *bag-of-features*) problem. These methods describe the input space as an orderless collection of local features and, recently, an accomplished impressive levels of performance has been reported [22]. The utilization of local features, which exhibit reliable properties to be detected and matched under different viewpoints, pose, or lighting conditions in the same scene, comprise a handy tool to solve recognition problems [23]. Due to the fact that these features abound in a scene, they should be stored in a representative manner, so as to produce appearance based descriptions [24,25,5]. Early works that utilize the *bag-of-features* problem to assist robotic navigation are those presented in [26,27]. The authors in [28,29] showed that effective methods for recognizing objects or places can be formed based on histogram-like descriptors. They proposed a method of handling higher-dimensional histograms, which have been utilized for a recognition task under the supervision of Gaussian derivative operators. The main advantage of such an attempt is that it produces appearance based information describing abstract representation of various aspects that different objects or places might exhibit. Additionally, histogram based methods for recognizing spatial events have been developed in [30,31], with an extension to local-spatiotemporal features. In a different approach, the authors in [32] made use of SIFT features to address the problem of building a higher level conceptual map directly from images. The main attribute of this method is that it produces topological maps by graph partitioning that describe spatial relationships between images.

When the conceptual representation of the input space is complete, the training of a system to robustly memorize the spatially variant place instances follows. Thus, a third way to distinguish the cognitive navigation algorithms relies on the classification algorithm they use. Several appearance based classification algorithms for place recognition are inspired by frameworks proposed for object recognition. For example, the work presented in [33] is a fully supervised SVM procedure and assumes that each place is represented by a collection of images during training, which capture the same visual appearance under different viewpoints and illumination variations. Moreover, the authors in [7] employed a combination of classifiers in a multi-modal framework for the perception of the environment. This system utilized various sensory inputs embedded in a high-level cue integration scheme, which relies upon SVM classifiers. In a different approach [34], the AdaBoost method is applied as a supervised learning algorithm to train a set of classifiers for place recognition based on laser range data. The authors in [35] developed a supervised learning approach to label different locations using higher level of cognition. They trained a classifier with visual features and then applied a Hidden Markov Model to increase the robustness of the final classification. In an experiment oriented work [6], the authors introduced a new dataset suitable for semantic navigation, inasmuch as it contains image sequences of several rooms under dynamic changes. They evaluated this dataset with an appearance based algorithm that integrated local features with SVMs making use of an ad-hoc kernel. To sum up, a conceptual comparison of several place classification approaches suitable for indoor robot navigation is provided in Table 1.

3. Algorithm description

Scholars in cognitive sciences argue that *space is at the heart of all conceptualization* [36]. Therefore, in order to augment robots' conceptualization, it is essential to construct a navigation paradigm which concurrently perceives geometrically accurate representations and provides semantic interpretations of the symbolic depictions. That said, in this section we propose a *cognitive navigation* framework based on a twofold architecture, which comprises two layers, namely the *numerical navigation* and the *semantic interpretation*. A semantically annotated topological graph coordinates these two layers passing the information from one layer to the other. This intermediate module is capable of augmenting the recognition and categorization capabilities of the integrated system, while it is able to distinguish between the places visited by making use of selected visual memories. The integrated framework is conceptualized in Fig. 2. The two layers including their intermediate coordination level is analytically described in the following subsections.

3.1. Low-layer navigation—3D SLAM

For the *low layer* navigation module, a SLAM algorithm with sensory input from an RGB-D camera is utilized. It is decomposed into five subordinate modules, as illustrated in Fig. 3. The first step comprises the detection and matching of SIFT features in successive color images. By evaluating the depth images of these

Table 1
Conceptual summarization of various place classification approaches.

Work	Recognition	Categorization	Semantic inference	Geometrical inference	Topology
[2]	✓		✓		✓
[6]	✓	✓			
[7]	✓	✓	✓		
[23]	✓	✓			
[29]	✓	✓	✓		
[33]	✓	✓	✓		
This work	✓	✓	✓	✓	✓

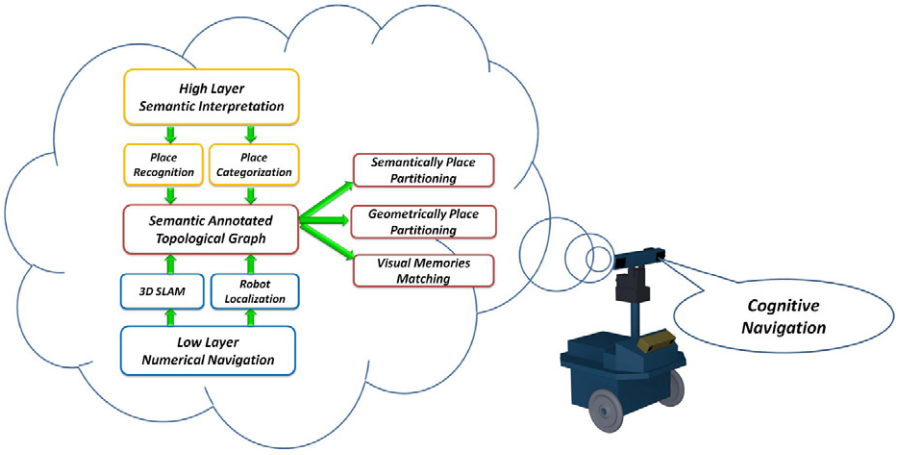


Fig. 2. Overview of the proposed framework.

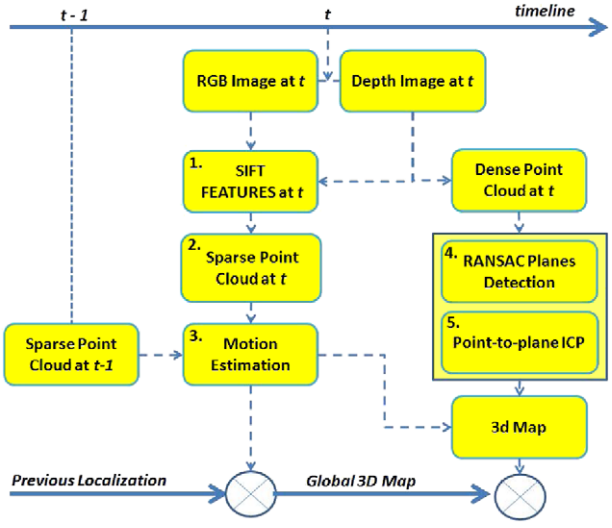


Fig. 3. Block diagram of the *low layer* navigation scheme.

feature points, a set of point-wise 3D correspondences between the consequent frames is obtained. As a matter of fact, two successive dense point clouds can be formed from the respective depth maps of the RGB-D sensor. Right afterwards, those 3D points are fed into a motion estimation routine, which nevertheless is not necessarily globally consistent, thereupon the result is refined in the fourth and fifth subordinate modules. In particular, two consecutive point clouds are first utilized and, by applying a simple RANSAC plane detection routine on both of them, only the subset of the 3D points that belong to the detected surfaces is retained. In the next step, these subsets of points are fed to an ICP algorithm to produce accurate refinements to the initial coarse motion estimation. The output of the algorithm at this stage is a globally consistent 3D model of the perceived environment, represented as a point cloud.

3.1.1. Motion estimation

Given the fact that our RGB-D sensor is fully calibrated and consistency among the active and passive cameras is ensured, the dimension of the resulted point cloud is an $M*N\times 3$ list of 3D points, where M and N are the rows and columns of the color images, respectively. However, for a rough motion estimation between the consecutive frames, only a subset of the initial point clouds is required. More analytically, among the potential detectors suitable for localization [37], the proposed motion estimation system employs SIFT, which detects and matches the salient 2D points within two consecutive frames. By evaluating the depth information at the respective 2D images, we obtain a set of point-wise 3D correspondences among the consequent frames. Let us assume that the robot observes a specific point ${}^{t}P$ in the 3D space, such as ${}^{t}P=[x_t, y_t, z_t]^T$. In the next time instance $t+1$ the robot undergoes a specific motion with rotation matrix ${}^{t}_{t+1}R$ and translation vector ${}^{t}_{t+1}T=[T_x, T_y, T_z]^T$, so the corresponding point ${}^{t}P$ is now observed as ${}^{t+1}P=[x_{t+1}, y_{t+1}, z_{t+1}]^T$. The transformation from point ${}^{t}P$ to ${}^{t+1}P$ is as follows [38]:

$$ {}^{t}P = {}^{t}_{t+1}R \cdot {}^{t+1}P + {}^{t}_{t+1}T. \tag{1} $$

The required rigid body transformation typically should conform with a sum of quadratic differences minimization criterion, resulting to a singular value decomposition (SVD) optimization problem, which in our case is equivalently handled by the Procrustes transformation [39] and results to the relative transformation of the robot. This way, a linear transformation is determined between the sparse 3D point clouds at time t and $t+1$ that corresponds to consecutive robot movements. However, this is a rough robot motion estimation, which frequently produces local minimum solutions. Therefore, an additional refinement step is essential to result in a more favoring solution.

3.1.2. RANSAC plane detection

Let us assume two dense point clouds ${}^{t}\mathbf{P}$ and ${}^{t+1}\mathbf{P}$ resulting from the depth images of two consecutive frames, respectively. The

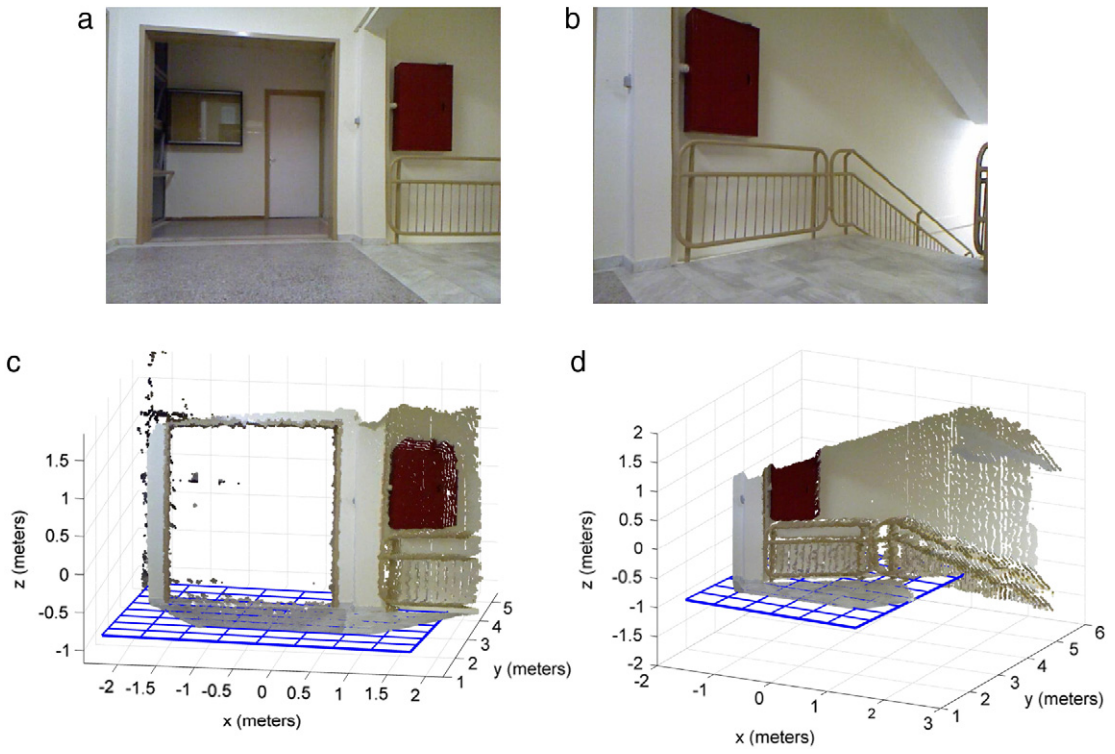


Fig. 4. (a) Color image at time t, (b) color image at time $t+1$, (c) textured 3D reconstruction by RANSAC plane detection at time t, (d) textured 3D reconstruction by RANSAC plane detection at time $t+1$.

attribute that pertains the two sets is that they share a respectable amount of common elements. This can be justified considering that the robots perform smooth movements between two successive time instances and the objects existing in the one frame do so, in majority, in the next one. Therefore, our model should rely on a specific constrain, i.e. the most prominent 3D planes that exist in the first frame should also appear in the second one. Consequently, two arbitrary subsets ${}^{t}\mathbf{P}' \subseteq {}^{t}\mathbf{P}$ and ${}^{t+1}\mathbf{P}' \subseteq {}^{t+1}\mathbf{P}$ can be found, such that ${}^{t}\mathbf{P}' = \{{}^{t}P'_i : i = 1, 2, \ldots, M\}$ and ${}^{t+1}\mathbf{P}' = \{{}^{t+1}P'_i : i = 1, 2, \ldots, N\}$, where ${}^{t}P'_i$ is a point belonging to cloud ${}^{t}\mathbf{P}$ at time t. These subsets of points retain a specific geometry and their further utilization in the calculation of the refined orientation of the robot, as well as in the generation of a consistent 3D map of the perceived environment, is analytically described Section 3.1.3. For the extraction of ${}^{t}\mathbf{P}'$ and ${}^{t+1}\mathbf{P}'$ we employ the RANSAC plane detection algorithm [40]. More specifically, three random points are selected to detect a plane by RANSAC, which in principle should provide adequate information about the estimation of the plane defined by them. The plane parameters are computed and then a score function is employed as a goodness-of-fit criterion. The respective score is adjusted by a threshold α, which specifies the number of points that belong to the plane. All the computed distances from any point to the detected plane should be less than α. The selected plane is the one achieving the greatest score [41]. In our case the reader should bear in mind that the designed methodology aims to be evaluated in indoor navigation scenarios, whereupon between two consecutive point clouds we only seek those subsets of points that belong either to the floor or to the walls of the room, i.e. the most prominent and large surfaces. For example, let us assume that we seek for a common surface into two consecutive frames, as depicted in Fig. 4(a) and (b). We perform the 3D reconstruction step and then we apply the RANSAC plane detection. As depicted in Fig. 4(c) and (d) the most prominent and common plane in the successive images is the one corresponding to the floor. Hence, we retain only those points that verify the equation of the plane due to the fact that they obey on the same geometry. Those features are then utilized for the refinement of the initial rough motion estimation. At this point it should be mentioned that multiple planes can be extracted from one scene by leaving out, on each iteration, the points that belong to detected planes and then repeating the RANSAC algorithm on the remaining 3D points of the initial point cloud.

3.1.3. 3D Map building using ICP

By applying the motion transformation on the respective 3D point clouds we obtain a rough alignment and, as a result, the 3D SLAM system suffers from erroneous registrations, as depicted in Fig. 5(e). By applying the motion transformation on the respective 3D point clouds we obtain a not that accurate alignment and, as a result, the 3D SLAM system suffers from erroneous registrations, as depicted in Fig. 5(e). Hence, following the RANSAC planes detection the 3D point clouds ${}^{t}\mathbf{P}'$ and ${}^{t+1}\mathbf{P}'$ are considered for the correction of the motion estimation. Indeed the set of points ${}^{t}\mathbf{P}'$ and ${}^{t+1}\mathbf{P}'$ now describe only the points of the subsequent detected planes at each time step, i.e. ${}^{t}[\mathbf{P}'_{s1}, \mathbf{P}'_{s2}, \ldots, \mathbf{P}'_{sn}] \subseteq {}^{t}\mathbf{P}'$, where $s1, s2, \ldots, sn$ denote the number of the detected planes. By considering the subset of points that fulfil the same criterion, i.e. they belong to the same geometric plane, we seek for a transformation that should correctly register the entire 3D point clouds among the consecutive time instances. The most commonly used algorithm to fine register the 3D point clouds is the Iterative Closest Point (ICP) one [42,43]. The novelty of the proposed work is that the ICP algorithm considers only the points that belong to specific geometric surfaces in consecutive time instances (e.g. floor and side-walls). The successive point clouds share a great amount of spatial proximity, due to the fact that a coarse alignment occurred during the motion estimation procedure. The benefit from this procedure is twofold: firstly we avoid multiple iterations restricting the rigid body transformation search by one order of magnitude in calculation time and, secondly,

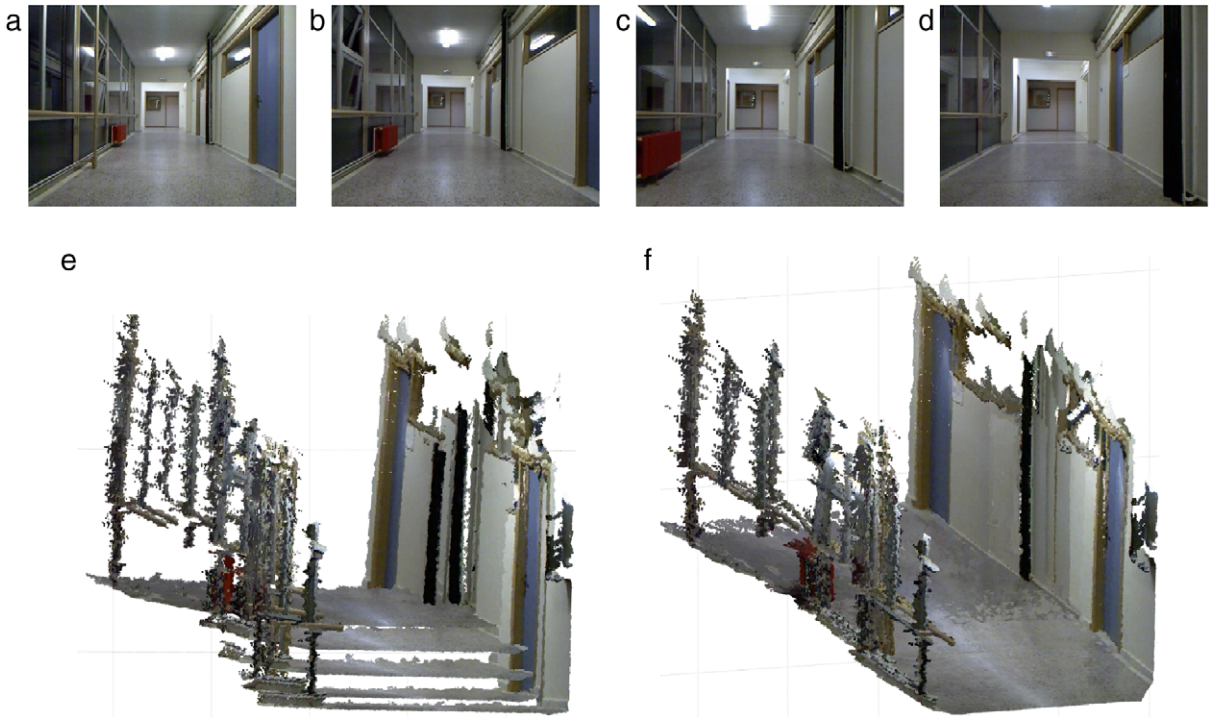


Fig. 5. (a) Color image at time t, (b) at time $t+10$, (c) at time $t+20$, (d) at time $t+30$, (e) the 3D map computed by applying the simple motion estimation transformation, (f) the 3D map computed by applying the RANSAC planed detection and point-to-plane ICP refinement steps.

we increase the likelihood to achieve an accurate solution. These advantages are feasible due to the fact that the considered points are contained in two successively observed scenes. Concerning the two successive 3D point clouds ${}^{t}\mathbf{P}'$ and ${}^{t+1}\mathbf{P}'$, we utilize a *point-to-plane* ICP algorithm [44], which seeks for a transformation K, that registers the two point clouds. The pseudocode of a simplified computation process calculating the best transformation K, given two consecutive point clouds, is described in Algorithm 1.

input : two point clouds: ${}^{t}\mathbf{P}'$, ${}^{t+1}\mathbf{P}'$
output: the transformation K that registers ${}^{t}\mathbf{P}'$ and ${}^{t+1}\mathbf{P}'$

$K \leftarrow K_0$;
while *no convergence achieved* **do**
  for $i = 1$ **to** N **do**
    $v_i \leftarrow Find\_The\_Closest\_Point\_in : {}^{t}P'(K \cdot {}^{t+1}p'_i)$;
    if $\|v_i - K \cdot {}^{t}p'_i)\| \leq th_{\max}$ **then**
      $bias_i \leftarrow 1$;
    else
      $bias_i \leftarrow 0$;
    end
    $K \leftarrow argmin\{\sum_i bias_i \|n_i \dot{(} K \cdot {}^{t+1}p'_i - v_i)\|\}$
  end
end

Algorithm 1: Point-to-Plane ICP algorithm, where $th_{\max}$ is a maximum distance threshold, v_i is the resulted transformation of each point and $bias_i$ is an indicator assigned to this point. The point-to-plane case minimizes the error along the surface normal, i.e. n_i is the surface normal at v_i .

The output of the point-to-plane ICP algorithm is a transformation $K = [{}^{t}_{t+1}R_{ICP}, {}^{t}_{t+1}T_{ICP}]$ that aligns the two successive point clouds. The transformation K is combined with the initial motion estimation and a refined estimation of the robot's pose is obtained, i.e. ${}^{t}_{t+1}R_{ref} = {}^{t}_{t+1}R \cdot {}^{t}_{t+1}R_{ICP}$ and ${}^{t}_{t+1}T_{ref} = {}^{t}_{t+1}T + {}^{t}_{t+1}T_{ICP}$. The calculated ${}^{t}_{t+1}R_{ref}$ and ${}^{t}_{t+1}T_{ref}$ are then involved to the accurate registration of the successive 3D point clouds. This procedure is performed separately in each time step and, hence, the 3D map of the explored area is constructed incrementally. Fig. 5(f) illustrates a short 3D map of a corridor, which corresponds to a 15 m route as a sampling of 40 successive robot observations. The resulted 3D map covers an area of approximately 30 m^2 with a resolution of 1 cm. Comparing Fig. 5(e), where the point clouds have been merged by applying only the initial motion estimation with the output of the 3D SLAM algorithm, as depicted in Fig. 5(f), one can directly realize that the second one produces more accurate and consistent results. That is, the aforementioned refinement invokes essentially on the 3D SLAM algorithm. Moreover, in the re-projection procedure, the walls, the floor and any of the objects appear mutually orthogonal and the scene is well reconstructed. The color values of the reference RGB-images have been quantized and re-projected to the 3D point clouds for illustration purposes.

3.2. High-layer navigation, semantic mapping

In this section we describe the *high layer* of the proposed cognitive navigation algorithm. As we have previously mentioned, one of the mandatory attributes that a robot should possess is to effectively produce semantic inferences irrespectively to its current location. While a mobile robot moves arbitrarily around within an indoor space, its sensors record an excessive amount of information, which corresponds to numerous observations. However, due to limited resources, the robot is able to memorize and recall only a finite number of representations of the explored space. In this work we propose a spatial abstraction of the input space for the efficient memorization of the distinct places (e.g. "office", "corridor", etc.). More specifically, from each sample of a queue of images, local descriptions of the most salient features are extracted and stored as raw information. These features describe the input space, i.e. any place that the robot should learn. Then a Neural Gas algorithm is employed to perform supervised space quantization. The latter has the capacity to quantize the input space by forming very abstract

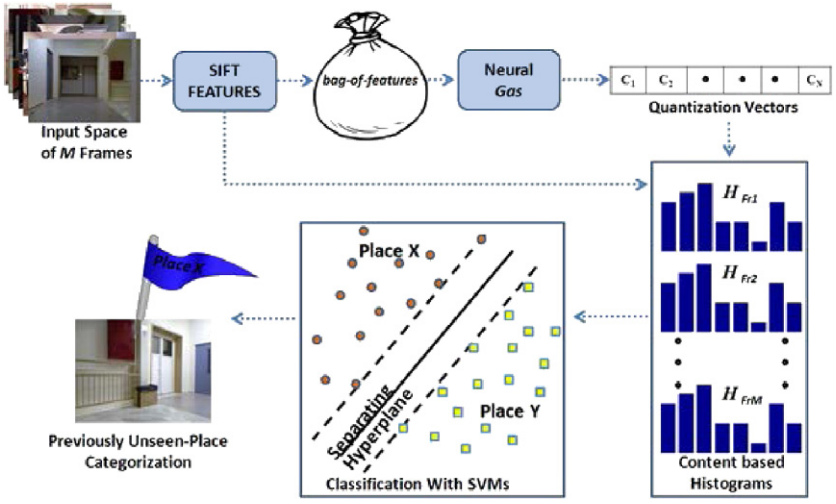


Fig. 6. Block diagram of the *high layer* navigation—semantic mapping.

descriptions able to surrogate the raw information. In the next step, consistent histograms for each sample of image sequence are created over the quantization descriptions and, as a result, the robot learns an abstract representation in each time instance. This procedure is shown in Fig. 6.

3.2.1. Bag-of-features problem

The SIFT detection and description algorithm computes salient points of a scene based on the appearance of the objects at particular interest points, whereas the resulted feature points are invariant to image scale and rotation [14]. Additionally, each detected feature in an image is registered with its location in image coordinates and a description vector, deriving by a histogram processing at the local neighborhood. The amount of features might vary from a few keypoints to many hundreds and it is regulated by a specific threshold. In the proposed work we selected a strict threshold, as it is a fair compromise to retain only the most salient features of the scene that exhibit repeatability in contiguous frames.

Let us assume that the robot should learn different places from a queue of M images that contain various representations of specific such ones. The SIFT algorithm is applied on every single image of the queue and the detected feature points are concatenated. The resulting feature space turns out to a *bag-of-features* problem that comprises a substantial description of the entire space to be memorized. The features space is denoted as a data matrix $\mathbf{S}^{128\times\sum_{k=1}^{M} n_k} = [\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_M]$, where each subordinate matrix $\mathbf{S}_k^{128\times n_k}$ corresponds to the output of the SIFT algorithm when applied on image $k = 1, 2, \ldots, M$ and n_k is the respective number of SIFT features. Moreover, each subordinated matrix is in turn defined as $\mathbf{S}_k^{128\times n_k} = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n]$, where each vector $\mathbf{s}_i$ corresponds to the description of a detected feature, such that $\mathbf{s}_i = [f_1, f_2, \ldots, f_{128}]^T$ and $i = 1, 2, \ldots, n$ indicates the number of the detected features in an image.

3.2.2. Appearance based histograms

Given the feature space $\mathbf{S}^{128\times\sum_{k=1}^{M} n_k}$ that describes all the places the robot should learn, an abstract representation should be formed so as to efficiently organize the redundant information. One almost straightforward procedure is to quantize the input space $\mathbf{S}^{128\times\sum_{k=1}^{M} n_k}$ in representative prototypes, i.e. quantization vectors. The space quantization – by employing clustering algorithms – constitutes a fundamental solution in various fields, such as pattern recognition, image processing, and machine learning [45]. In this work we have adopted the Neural Gas algorithm [46,47], which is an artificial neural network and its basic objective is to optimize a cost function which minimizes the quantization error. More analytically, given the substantial description of the space $\mathbf{S}^{128\times\sum_{k=1}^{M} n_k}$ (*bag-of-features*), a subset of representative vectors should be defined that characterize the entire input space in an abstract form. Therefore, a set of Q quantization vectors $\mathbf{C}^{128\times Q} = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_Q]$ is now sought, which provides a satisfactory representation of the initial space, insofar as an error e is not outdone. Moreover, the Neural Gas algorithm returns an index $w_j, j = 1, 2, \ldots, Q$ for each feature of the space $\mathbf{S}^{128\times\sum_{k=1}^{M} n_k}$, that denotes the identity of the quantization vector in which the specific feature is registered. The maximum number of epochs $t_{\max}$, required for the Neural Gas to produce its output, is a user specified parameter. Typically, the number of the quantization vectors Q and the specified epochs of the convergence procedure $t_{\max}$ affect the resulting representation error of the model, i.e. the larger the number of the of quantization vectors Q the more accurate the space reproduction performance. These quantization vectors correspond to the "vocabulary", which contains all the visual words describing the initial space in an abstract form.

In the next step, the quantization vectors (visual words) are employed to form the appearance based histograms that represent the initial input images. Given the previously calculated matrix of features $\mathbf{S}_k^{128\times n}$ it is feasible to form a representative consistency histogram $h_{\mathbf{S_k}} \in \Re^Q$ for each image $k = 1, 2, \ldots, M$ spread over the Q quantization vectors. Applying this procedure on the entire queue of images, a data matrix $\mathbf{D}^{128\times M} = [h_{\mathbf{S_1}}, h_{\mathbf{S_2}}, \ldots, h_{\mathbf{S_M}}]$ is created and used for the training of an SVM. More analytically, the L2 norm between the detected features and the quantization vectors is calculated resulting to the formation of the representative histogram; the binning is performed according to the smaller distance. Consequently, each image in the sequence has been replaced by a respective appearance based histogram, which constitutes the respective feature vector to train a classifier. Moreover, it is worth noting that the "vocabulary" is created only once in an off-line procedure from a queue of images describing all the places that should be memorized (e.g. "office", "corridor", etc.) and, therefore, it does not evolve with time. The latter means that each query image has to be represented by a histogram circumscribing the scene with respect to the already existing knowledge. Therefore, the contribution of these histograms is twofold: firstly the initial space has been replaced by appearance based representations able to capture the geometrical attributes and the semantic content of the scene, and

secondly the input space is described in an abstract yet consistent manner. It should be mentioned that the computation of the "vocabulary" is a time consuming procedure, nevertheless the Neural Gas quantization vectors are calculated and stored only once, during an off-line procedure. The reposited quantization vectors are utilized to perform on-line testing, which comprises only the execution of the SIFT descriptor and the formation of the histogram for each image frame by calculating the L2 norm and performing the binning procedure.

With aim to make our "vocabulary" method more distinguished we compare the proposed method to the "vocabulary trees" appeared in [48], where the visual words correspond to the detected features. In this case the Maximally Stable Extremal Regions (MSERs) detector has been utilized as a variation of the SIFT algorithm and, then, the visual words are organized in a multilevel hierarchical structure, which corresponds to a "vocabulary tree". This structure allows fast image retrieval from massive populated collections, utilizing either the L1 or the L2 norm optimization criterion. Moreover, in [48] it has been shown that the performance is greatly influenced by the size of the retained database; thus the recognition accuracy – in the range of 10 K to 1 M of images – varies from 76% to 70% respectively. While in the aforementioned procedure the "vocabulary" corresponds to the detected features, in our case the resulted "vocabulary" stands for the calculated representative prototypes, as a result of the space quantization procedure of the Neural Gas. The formation of each histogram comprises a mapping of the query image in the "vocabulary" space. Therefore, the bins of the histogram that will be filled depends on the current place of the robot. Owing to this fact, the proposed system can utilize the same "vocabulary" for both recognition and categorization, achieving remarkable generalization capabilities (Section 5). Additionally, a more detailed "vocabulary" will result in more consistent appearance based histograms during the inference mode. Last it ought to be noted that comparing with the work described in [48] the amount of the retained prior knowledge to the system is deteriorated, indicating enhanced performance due to the fact that the resulted histograms are built respecting to the co-occurrences of the learned space and the target one.

3.2.3. Place classification

A robust place classification framework should comprise both recognition and categorization capabilities. On the one hand, place recognition refers to the robot's competence to efficiently recognize a place, given that it has been previously trained on similar scenes. One the other hand, the place categorization is a challenging task, as it comprises the correct place recognition of the robot without being aware of any prior knowledge. The latter means that, while the system is trained on a specific dataset comprising images of places in a specific building, it is tested on similar places that belong to a different dataset containing places of another building. The main objective of the proposed work is to prove that the appearance based histograms, introduced here for the first time, hold great descriptive, discriminative and generalization capabilities and can be used to efficiently solve the classification tasks within a cognitive navigation framework. This statement is further supported by the adoption of a linear classifier, namely an SVM [49], the performance of which greatly depends on the quality of the utilized feature vectors. The only regulation parameter for this classifier is the C one, which penalizes the large errors. Howbeit, in this work a small value has been set to this parameter, exhibiting the increased confidence the authors have in the formed feature vectors. A more complex kernel function, such as the polynomial or the Gaussian one, would not enhance further the ability of the proposed appearance based histograms to capture a consistent spatial representation of the initial input space. For the implementation of the SVM the LIBSVM library was used [50]. The SVM is by its nature a binary classifier; however, its extension to multiclass problems can be performed in several ways [51]. In this work the one-against-all strategy has been preferred, that for each different class a respective SVM is trained, to separate this single class from all the others.

3.3. Semantically annotated topological graph

Along the robot's wandering a semantically annotated topological graph is formed. Each of its nodes is added to the graph with respect to two different criteria. The first one relies on geometrical constrains and, consequently, the robot should drop a node in the graph, after it has traveled a certain distance. This distance is calculated utilizing the visual odometry output of the *low layer* navigation scheme. The robot's location is estimated in each step and a new node is added to the graph when the current position of the robot and the last added node exceed a specific distance d. The threshold d in this paper equals one meter, above which covers the situations that the velocity of the robot is such that the scene might change significantly. The second criterion relies on the decision of the *high layer* module according to the semantic inference made for the current place. More specifically, a voting procedure of the SVM models decides which place label should be assigned to the respective node on the topological graph. The w neighbors of the query image participate in a majority vote process. This constrain is associated with the time proximity of the successive frames during the robot's exploration and boosts its confidence for the place categorization. The edges of the graph denote the Euclidean distance between the nodes.

The main attribute of the constructed topological graph is that during its formation, characteristics from both the high and low layers are exploited, bringing together geometrical and semantic information. The appearance based histograms that correspond to the detected nodes constitute visual memories to be utilized both in the place categorization and in the detection of multiple places with the same label. More analytically, the semantic topological graph is constrained by the currently estimated location of the robot. The detection of multiple places e.g. "office1" and "office2", that bear the same label is accomplished by applying simple checks expressed as their in between L2 norm. Actually, considering the topological graph in Fig. 7 it is revealed that the each place is described by nodes sharing the same semantic label retaining spatial proximity. Therefore, when the robot moves in an indoor environment each time a node is pooled in the topological graph, both its spatial coordinates and the semantic label are checked. In the specific cases where the robot drops a new node of the same label with an already existing place e.g. "office", then all the retained visual memories that possess the same label are compared. In case that the new node does not resemble – in terms of Euclidean distance – with the already pooled ones, then there is an indication that the robot is located in different place (e.g. another "office"). Note that in this case we set a very strict threshold (0.8) during the comparison of the appearance based histograms, in order to capture the differentiations of the rooms that belong to the same category. In the next step, the new candidate node is further compared with the existing nodes considering their geometrical distribution within the explored space. If the distance of these nodes is greater than a predefined threshold (e.g. 5 m), then the newly added node is considered to be a different place of the same type i.e. "office2" as depicted in Fig. 7. However, in the situations where the robot returns to an already visited place, then the aforementioned geometrical constrain prohibits the formation of a different place of the same type and, therefore, the new node on the graph retains the label of the already explored place.

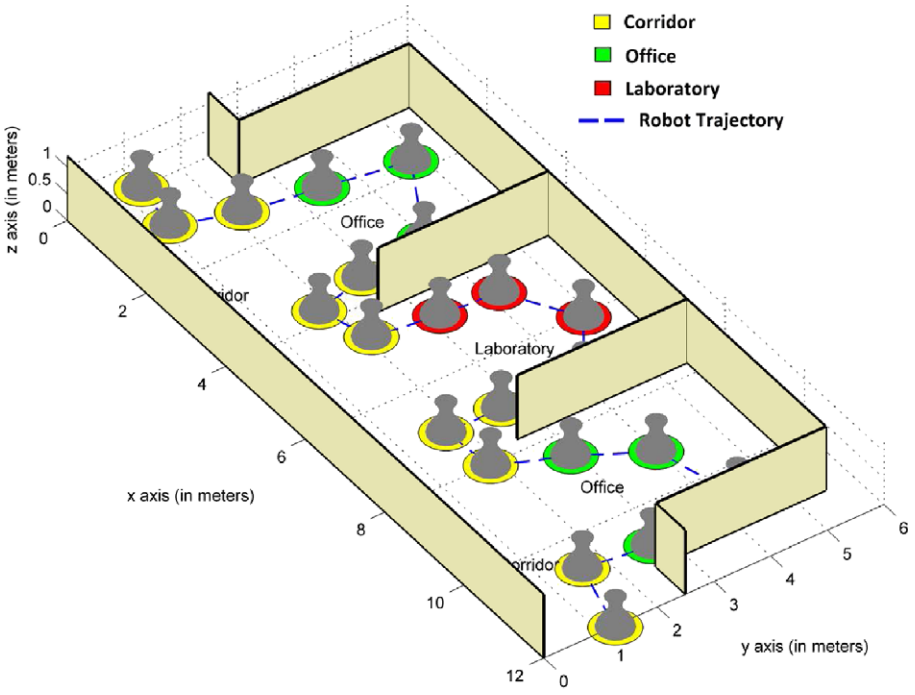


Fig. 7. An example of the semantically annotated topological graph. The latter is used to combine the geometrical with the semantic information of the *low* and *high layer* respectively to differentiate among distinct places of the same type. In this example there are two "offices" with Euclidean distance more than 5 m, while the visual memories of the first mapped "office" does not resemble closely with the new candidate node. Consequently, the exploitation of the topological map resulted to the partition of the second "office" as a different place.

4. The Cognitive Navigation dataset

An additional goal of this work is the introduction of the Cognitive Navigation Dataset, intended for use in indoor localization and semantic mapping. This dataset has been captured with a Kinect RGB-D sensor mounted on the MAGGIE robot, 0.8 m above the ground and tilted by 2.54° and it consists of three different parts:

- Part A involves images of a corridor, an office, a laboratory, a bathroom and an elevator area, captured in natural lighting conditions (daytime light). This dataset segment contains 1286 images including the respective depth maps.
- Part B involves exactly the same places as in Part A, but they have been captured during nighttime and, therefore, the illumination conditions are artificial. This part contains 1292 images including the respective depth maps.
- Part C has also been acquired under artificial lighting conditions and comprises a continuous exploration of all the aforementioned places. The movements of the robot are very smooth preserving similar field of view in successive frames. This dataset section contains a continuous sequence of 557 images and corresponds to 70 m route.

The first two parts of the dataset are suitable for the validation of place classification algorithms, i.e. involving recognition and categorization only, due to the fact that it includes images from places with different viewpoints captured under variant lighting conditions. The third part of the dataset, apart from recognition and categorization, is also suitable for indoor localization and mapping, due to the fact that the consecutive frames share great spatial and time proximities. In Fig. 8 some examples of the captured images are depicted. The Cognitive Navigation dataset is freely available to the research community and can be retrieved, along with further technical information, at [52].

5. Experimental results

The proposed navigation framework has been evaluated on two different datasets. The first one is the Cognitive Navigation dataset introduced in the previous section and the second one is the COLD [6]. The latter is a large database suitable for vision based recognition systems, inasmuch as it consists of three subdatasets, acquired in different Universities. In this work, the subdataset captured at the Visual Cognitive Systems Laboratory at the University of Ljubljana has been utilized, due to the fact that the places contained share semantic consistency with those presented in the proposed Cognitive Navigation dataset. An additional attribute that makes this set of images suitable for the evaluation of place recognition algorithms is the fact that it contains scenes captured under different lighting conditions i.e. "sunny", "cloudy", and "night". The evaluation of the proposed navigation framework has been performed in four different phases. The first one comprises the evaluation of the *low layer* navigation framework on the Cognitive Navigation dataset. The second one proves the place recognition capacity of the proposed method, which is evaluated in both datasets. The third one involves the assessment of the place categorization capabilities of the introduced method, given the fact that the system has no prior knowledge of the places where it is exposed during the testing procedure. The fourth phase appraises the ability of the integrated system to draw semantic inferences while the robot builds a 3D map of the explored area. The place classification capabilities of the introduced framework are compared with the results of the method presented in [33], which has also been evaluated with the COLD dataset.

The place classification algorithm, which is proposed here, comprises the regulation of two parameters only. The most important is the selection of the number of the space quantization vectors in the Neural Gas algorithm. The ability of the quantization vectors to represent the initial input space determines its impact. As it is

Fig. 8. Image samples for (a) Part A, comprising samples shot under natural lighting conditions, (b) Part B, comprising samples shot under artificial lighting conditions and (c) Part C, a certain root of the robot in artificial illumination conditions.

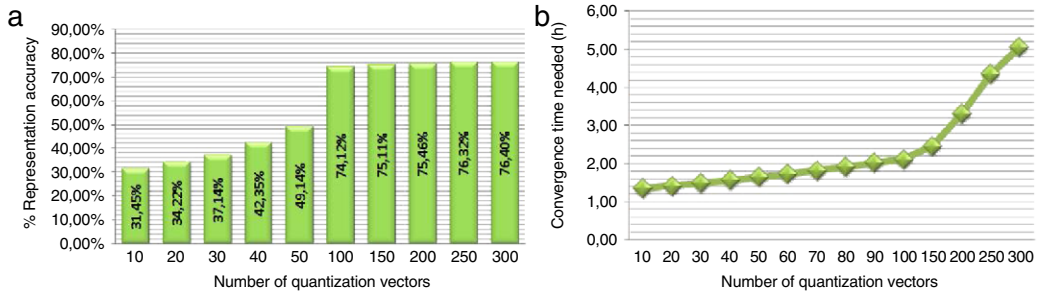


Fig. 9. (a) The space representation accuracy against the number of quantization vectors and (b) the required convergence time for various populations of quantization vectors.

depicted in Fig. 9(a), the representation accuracy increases almost linearly for the first 100 vectors, but then it saturates in the interval [100, 300], where the representation accuracy remains constant about 75%. Moreover, it should be mentioned that, while the number of the quantization vectors increases, the required convergence time increases accordingly, as it is illustrated in Fig. 9(b). Therefore, the selection of 100 quantization vectors, that correspond to 74.65% reproduction ability of the initial input space, is a decent compromise between the accuracy and the required convergence time. The same value has been retained during the entire experimental procedure ensuring uniformity and comparability among the results. The kernel type of the SVM and its internal parameters should also be selected, which in our case is a linear kernel SVM classifier. The selection of such a weak kernel relies on the wish to demonstrate the capabilities of the appearance based histograms to capture the most representative characteristics of the input space. The regularization parameter C was set equal to 10.

3*D SLAM*. The *low layer* navigation algorithm being the cornerstone for a successful navigation framework has been firstly evaluated. The proposed 3D mapping system has been applied on the Part C of the Cognitive Navigation dataset and the results of the proposed algorithm are shown in Fig. 10, where a top down projection of the formed 3D map is illustrated. In this figure, one can observe that the proposed method provides a perspective environment mapping with no accumulative errors, proving that the RANSAC plane extraction combined with the point-to-plane ICP registration concludes in a consistent 3D map of the indoors environment. As a result, the entire structure of the explored places is well reconstructed exhibiting 1 cm accuracy, i.e. insofar as the accuracy the involved RGB-D sensor permits. However, the creation of a 3D map with great consistency is not the sole objective of this paper. Actually, it aims to introduce a functional *low layer* navigation algorithm, which permits interaction with the *high layer* semantic one to form an holistic cognitive navigation framework that affords the robot full position information about itself, both numeric and semantic.

Place recognition. Firstly, the recognition ability of the proposed method for the COLD dataset is examined. This dataset consists of six different places, i.e. a 2-persons-office, a corridor, a laboratory, a printer-area, an elevator and a bathroom shot under different lighting conditions. Several experiments have been performed to assess the classification capacity of the proposed algorithm. The first experiment comprises the training and testing of the classifier using the one-against-all procedure for each of the lighting conditions separately. In order to ensure the statistical significance of the results and to avoid over-fitting models at the same time, the 10-fold cross validation procedure has been adopted. In Fig. 11 the averaged results are presented along with their respective standard deviations to summarize the recognition accuracy for all the six places under unlike lighting conditions. The depicted average recognition accuracy is considerably high and varies between 94% and 99%, retaining narrow standard deviations, as previously.

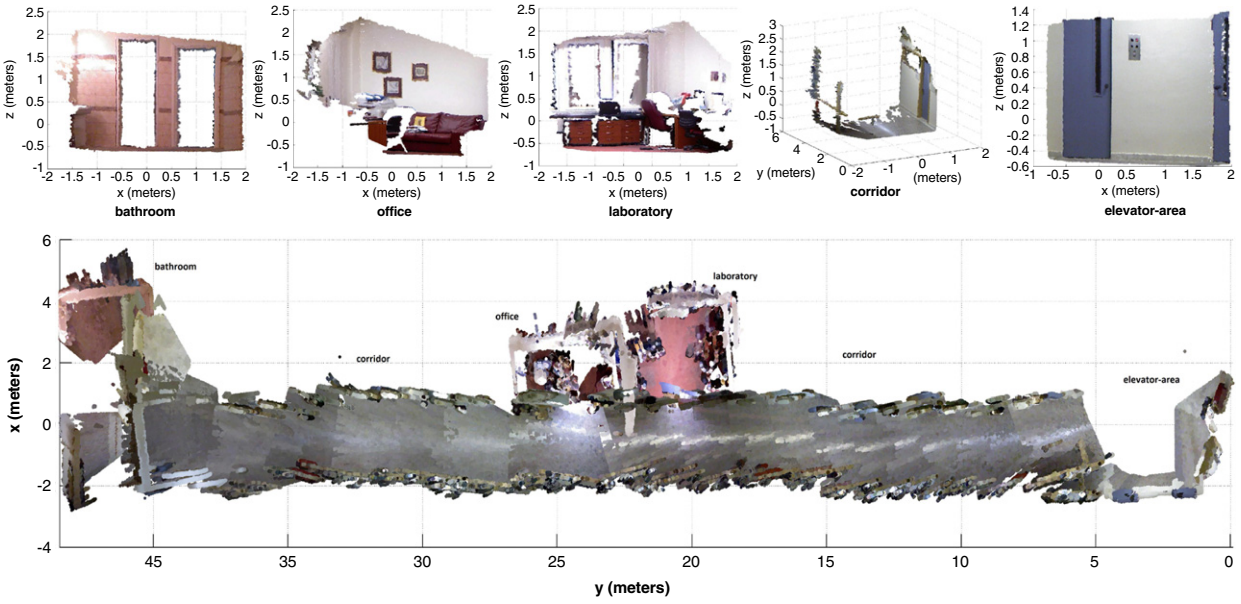


Fig. 10. The first row depicts the 3D reconstruction of five instances corresponding to the five distinct places of the dataset, while the second row illustrates the 3D global map along the 70 m path.

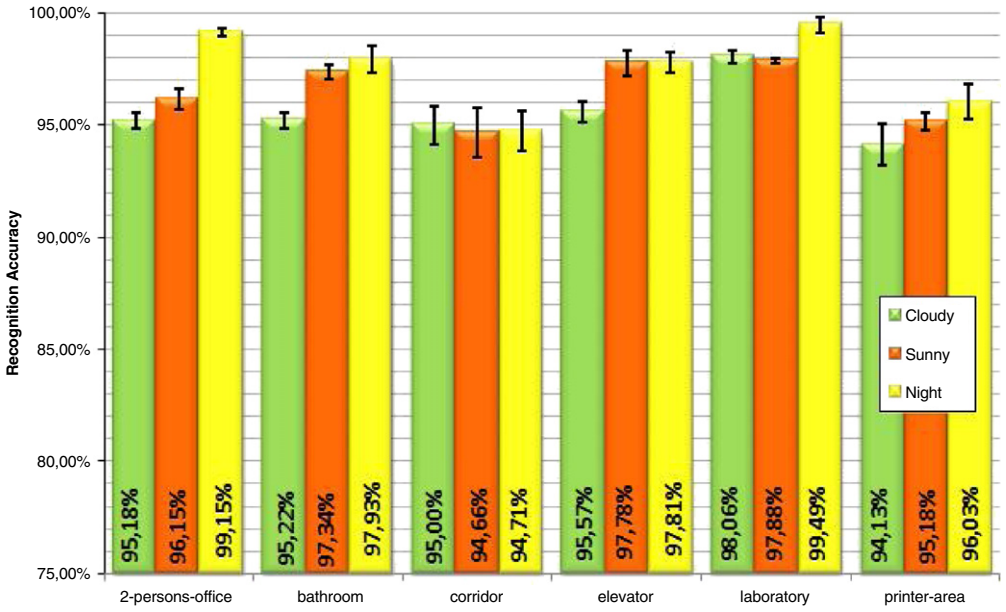


Fig. 11. The averaged recognition accuracy of the proposed method under variant illumination conditions for the multiple places of the COLD dataset. In each bar the averaged accuracy along with the respective standard deviation is annotated.

The accuracy generally tends to decrease at the cloudy section of the dataset, due to the fact that many images were textureless and produced almost no SIFT features. Moreover, the "corridor" is the class that achieved the worst recognition percentage, as the input images exhibit increased otherness form each other; thus some instances have been considered in total different classes. The rest of the classes exhibit great recognition accuracy indicating the robustness of the appearance based histograms. The recognition capabilities of the proposed method have also been tested for the proposed Cognitive Navigation dataset. The "natural illumination conditions" correspond to the "sunny-scenario" and the "artificial illumination conditions" to the "night-scenario" of the COLD dataset. Once again, this experiment comprises a 10-fold cross validation utilizing the one-against-all procedure under each illumination condition separately. The results are illustrated in Fig. 12, where the proposed place recognition algorithm achieves also great recognition accuracy. The latter varies between 95% and 99%, with very tight standard deviations, indicating their statistical significance. Contrary to the COLD dataset the "corridor" presents great uniformity and, therefore, the accuracy is improved. However, the classification accuracy on the places "office" and "laboratory" is decreased, due to the fact that the objects in these two rooms exhibit exactly the same texture. Yet, the performance of the proposed approach remains excessive, i.e. over 95% recognition accuracy has been achieved.

A different and more challenging recognition task involves the evaluation of the proposed method throughout all the datasets available. Particularly, the proposed method has been trained with all the different classes under any illumination condition. This experiment reveals the ability of the formed content based

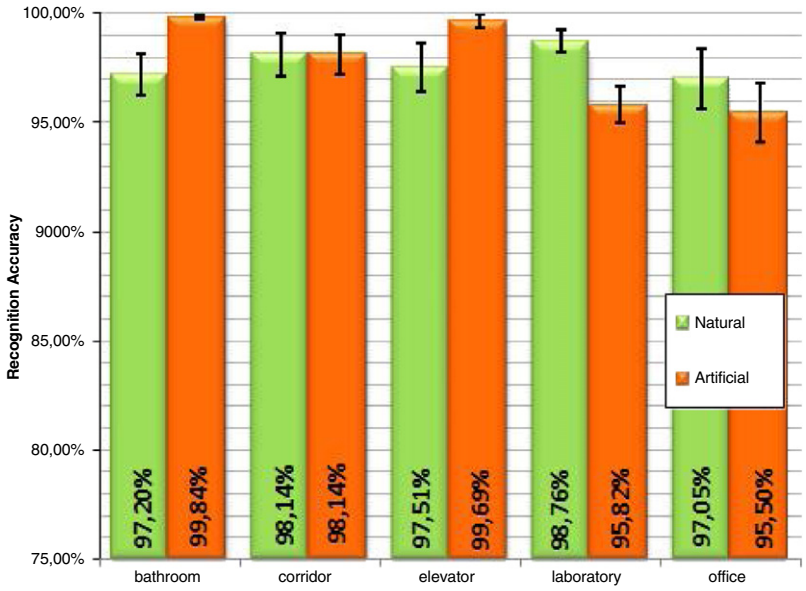
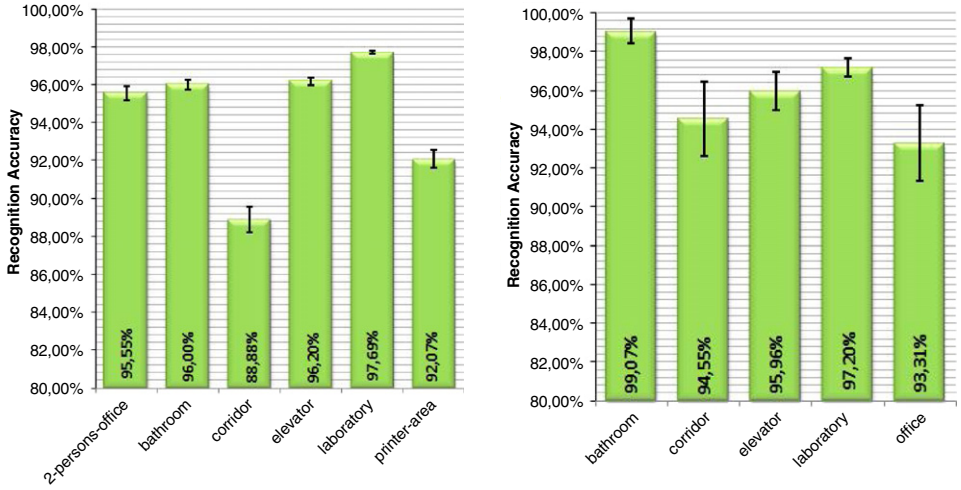


Fig. 12. The averaged recognition accuracy of the proposed method under variant illumination conditions for the different places of the Cognitive Navigation dataset. In each bar the averaged accuracy along with the respective standard deviation is annotated.



(a) COLD dataset. (b) Cognitive Navigation dataset.

Fig. 13. The average recognition accuracy under fused illumination conditions for the (a) the COLD dataset and (b) the Cognitive Navigation dataset. In each bar the averaged accuracy along with the respective standard deviation is annotated.

histograms to capture the geometric and texture characteristics of the input space under great variance in the illumination conditions, which is a heavy recognition problem. In order to track the performance of the proposed algorithm, the parameters both for the Neural Gas and the SVM remained the same, as well as the setup of the experiment remained intact. In this procedure the 10-fold cross validation for the training and testing has also been adopted; the results were averaged and presented along with their respective standard deviations. The results are summarized in Fig. 13(a) for the COLD dataset and in Fig. 13(b) for the Cognitive Navigation dataset, respectively. Considering the results in the Fig. 13(a), the proposed method exhibits significant recognition accuracy, which lies between the interval 88% and 97%. Once again, the algorithm achieved relative low performance in the class "corridor", which normally contains great non-uniformity and becomes more uneven under different illumination conditions. The proposed method exhibits also important performance in the Cognitive Navigation dataset and the recorded accuracy is within the interval 93% and 99%. One significant observation is that the algorithm constantly achieves significant performance in the recognition of the place "bathroom"; thus the appearance based histograms have captured its unique characteristics (texture and geometry of the existing objects).

Place categorization. An additional objective of the proposed work is to provide a robust place categorization solution. The latter stands for the ability of the robot to provide correct semantic deductions about places that has never visited before. For example, once the system has learned to distinguish the class "corridor" from the class "office" in the premises of the University of Ljubljana, then it should be able to distinguish akin places in any other building. The class "corridor", contains characteristics that are global for any corridor, such as windows, side walls and doors placed in a specific topology. The same observation also stands for the class "office", which normally consists of computers, chairs, etc. The ability of the robot to infer about places that has no prior knowledge of, i.e. to be trained on the COLD dataset and tested on

(a) Cognitive Navigation dataset.

(b) COLD dataset.

Fig. 14. (a) The places in the Cognitive Navigation dataset, (b) the places in the COLD dataset.

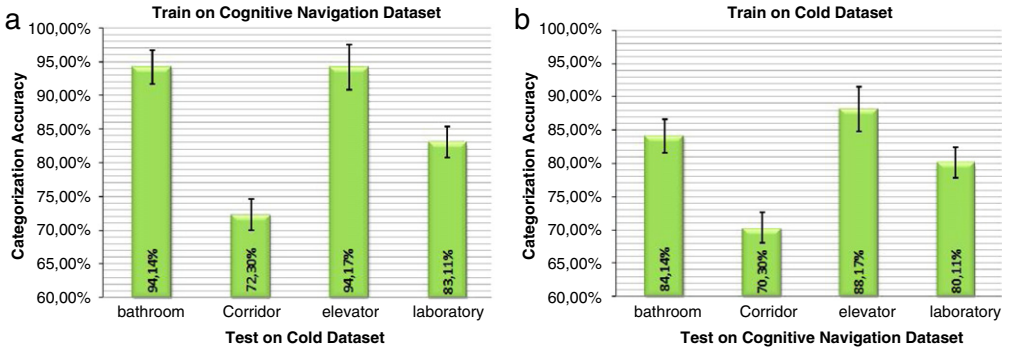


Fig. 15. The average place categorization accuracy utilizing as a training test (a) the Cognitive Navigation Dataset and (b) the COLD dataset. In each bar the averaged accuracy along with the respective standard deviation are annotated.

the Cognitive Navigation dataset, reveals the strength of the system to create consistent semantic based representations of the explored place, that possess ample generalization capabilities. This experiment took place into two distinctive parts: firstly the system has been trained on the Cognitive Navigation dataset and then tested on the COLD dataset, and vice versa. The total process ensures that the proposed algorithm has the ability to learn the spatial representation of the explored places independently to the utilized dataset. The method has been evaluated for the common classes among the two different datasets, viz. the "corridor", the "laboratory", the "bathroom" and the "elevator-area". Fig. 14 presents instances of the two datasets for the corresponding places, exhibiting the non-uniformity of the data. In this experiment, the one-against-all training and testing procedure has also been utilized. However, the adoption of the k-fold cross validation is meaningless in this case, due to the fact that the system should be tested on previously unseen data and, consequently, a specific validation procedure has been utilized. Let us assume that the COLD dataset is used for the training of the system. The respective training data is a selection following a random permutation procedure on the entire train dataset (in this case the COLD one). The same concept is followed in the testing dataset, which is the Cognitive Navigation dataset and, after the repetition of this routine for 100 times, the results are averaged and the standard deviation for each class is computed as depicted in Fig. 15. By following this multiple training and testing procedure, it is ensured that typical problems such as the model overfilling are avoided and the real categorization capabilities of the examined method are revealed. In particular, Fig. 15(a) describes the occasion where the system has been trained on the Cognitive Navigation dataset and tested on the COLD dataset. The system exhibits great categorization accuracy, i.e. 72%–94% proving that the appearance based histograms are able to deliver apt generalization capabilities. The significance of the results becomes bolder due to the unevenness of the two different set of images, as observed in Fig. 15(b). Additionally, it should be stated that the categorization capabilities of the proposed cognitive navigation system retain statistical significance by examining the standard deviations of the averaged accuracy, over the different places. A similar experimental procedure is presented in [6], where the same data are utilized for the evaluation of the place classification capabilities of an appearance based approach [33]. Therefore, an additional experiment has been conducted comparing our results with the classification accuracy as described in [6]. Following the concept of this work, the data in the COLD-Ljubljana subdataset have been split according to the different illumination conditions, e.g. train on the "cloudy" data and test on the "sunny" and "night" data. The results have been averaged among all the classes and the different types of data in terms of illumination conditions. Table 2 summarizes the classification results for the proposed method comparing to the one described in [33]. The appearance based histograms retain great recognition capabilities by capturing competently the salient characteristics of the scenes in each case, exhibiting overall recognition accuracy better that 88%. It should also be mentioned that in both the examined methods the training

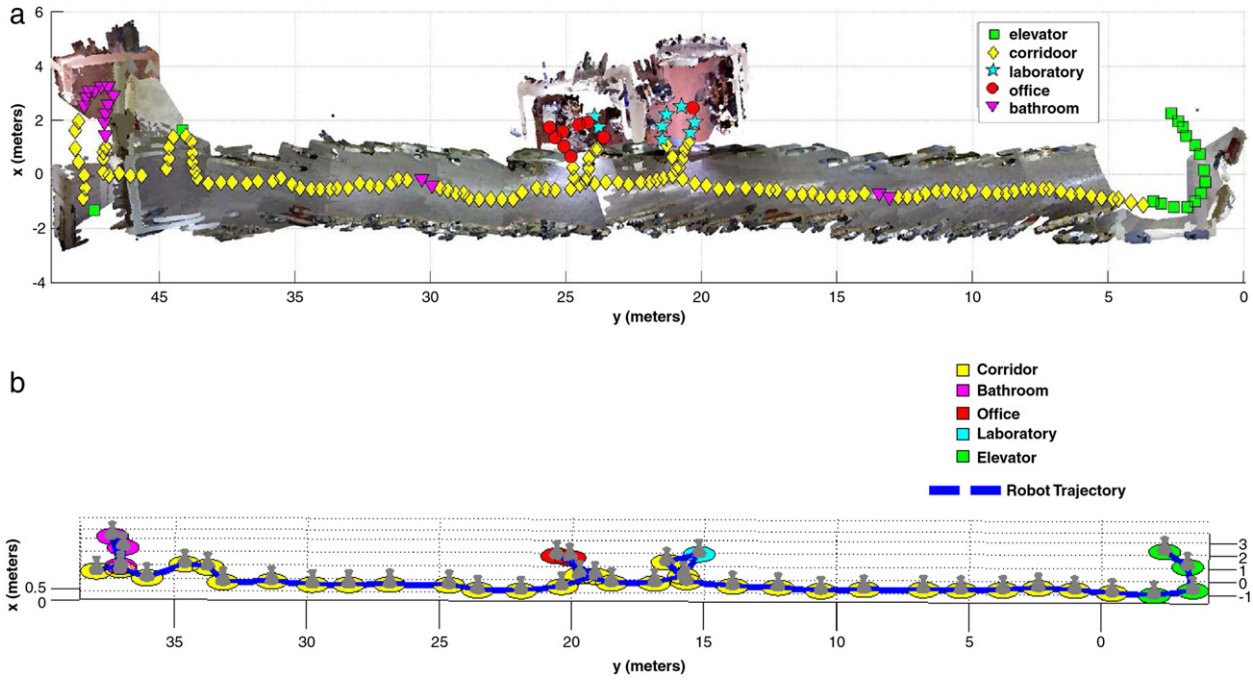


Fig. 16. (a) The recognition labels overlying the 3D global consistent map. Note that only a subset of the category labels is illustrated in this figure, as a result of a sub-sampling every 0.5 m. (b) The respective semantically annotated topological graph of the explored environment. The detected places have been formed appropriately and comparing to the figure above, the miss-classified instances have been filtering as a result of the voting procedure.

Table 2
The average classification accuracy for the fused datasets including all the existed classes in the COLD-Ljubljana sub-dataset. Note that the diagonals indicate that the training and testing on the same dataset result to better classification capabilities; therefore the categorization task is more challenging than the recognition one.

Train	Method described in [33]			Proposed method		
	Test			Test		
	Cloudy	Night	Sunny	Cloudy	Night	Sunny
Cloudy	80.48%	77.77%	79.74%	94.20%	91.43%	92.82%
Night	73.70%	91.86%	71.61%	89.58%	95.54%	88.34%
Sunny	79.82%	74.95%	84.20%	90.32%	91.64%	96.56%

and testing on the dataset with the same illumination conditions exhibit better performance than in the case where the system is tested on different illumination conditions. The latter proves that the categorization task is more challenging than the recognition one. However, the proposed method exhibits competent performance both in the recognition and the categorization retaining robustness in the semantic interpretation of the *high layer* navigation algorithm as a part of a complete cognitive navigation framework.

Place labeling. The third step of the experimental evaluation comprises the ability of the system to produce semantic inferences while the robot builds a 3D map of the explored area. Consequently, an additional experiment has been performed, where the place recognition capacity of the *high layer* framework is examined. The prerequisite to this step is that the robot should construct a consistent 3D map as it moves from one place to another. During this experiment the system was trained with images from the entire Cognitive Navigation dataset including both natural and artificial illumination conditions. The testing procedure constitutes an integrated framework, where the system constructs the 3D map of the visited places and simultaneously makes semantic inferences concerning the different places. In Fig. 16(a) the output of this procedure is depicted with the place labels overlying the 3D map. Specifically, the achieved place recognition accuracy is comparable to the one shown in Fig. 13(a), where the system was trained for and tested in the entire Cognitive Navigation dataset. At this point we should mention that the erroneous place labeling usually appears in course of the transition between two rooms, where both places are visible by the robot. Moreover, from Fig. 16(a), additional observations could be obtained about the performance of the place labeling algorithm; the laboratory is usually confused with the office, while the corridor is erroneously recognized as being the bathroom. However, in this experiment the system recognized correctly the 540 out of the 557 instances resulted in a 96.95% classification accuracy. Fig. 16(b) depicts the respective semantically annotated topological graph of the explored environment. As it is clearly illustrated in this figure the nodes are not uniformly dropped in the graph as a result of the two different criteria that govern its formation. Note that erroneous recognitions have been filtered out during the voting procedure of the SVM models, while the formed nodes indicate the correct place partitioning and recognition of the algorithm. It should also be mentioned that in the "laboratory" some nodes in the topological graph have been annotated as "corridor" due to fact that the each node in the topological graph is placed at the base of the robot irrespectively to what appears to its sensory input. More precisely, during the robot's travel the RGB-D sensor perceived more information from the "corridor" rather that the "laboratory", notwithstanding it was deploying in the "laboratory".

6. Conclusions

In this work a two-layer cognitive navigation algorithm has been introduced, motivated by the wish to concurrently accomplish three different objectives. The first one refers a *low layer* navigation, which comprises the accurate localization of the robot and the simultaneous formation of a consistent map of the explored area. As a response to this problem, the proposed work introduces an algorithm producing SLAM from RGB-D data. The novelty of the presented method is that in each consecutive frame the proposed system performs a RANSAC 3D plane extraction step retaining only the points that belong to the detected planes. Taking advantage of the specific geometrical topology of these points a point-to-plane

ICP algorithm is employed to merge the entire point clouds. This procedure produces refinements to the initial coarse motion estimation, which is performed by exploiting a SIFT feature tracking, while the outcome is a consistent 3D map of the places that the robot has visited. The *high layer* algorithm encapsulates all the cognitive attributes that a robot should possess to effectively draw semantic inferences irrespectively to its current location. The presented work deals with this objective by applying solutions for both the place recognition and navigation tasks. Moreover, the novelty at this point is that the input space is described by content based histograms, which endow the classification system with great descriptive, discriminative and generalization capabilities. The theoretically infinite input space is treated as a *bag-of-features* problem – formed by SIFT features – and the Neural Gas algorithm is employed to learn an abstract representation of the space. Each place instance is represented by an appearance based histogram allowing the system to retain short, yet meaningful descriptions. The classification method employs linear SVMs and exhibits great accuracy revealing the robustness of the content based histograms. The recognition and categorization capacities of the proposed cognitive navigation framework have been thoroughly examined in two different datasets and exhibited remarkable accuracy. The first dataset – introduced in this paper – is the Cognitive Navigation dataset, which is suitable for place classification and localization algorithms and the second one is the COLD. The proposed method exhibits noteworthy recognition capabilities achieving over 95% accuracy in all cases. Additionally, the system exhibited significant place categorization capabilities achieving over 70% accuracy assessed in previously unseen data. The reported method exceeds other similar methods in classification and categorization efficiency, as evinced by comparative a study conducted.

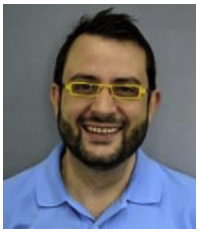
In conclusion, the proposed methodology constitutes an integrated framework for cognitive robot navigation endowing artificial agents with accurate position information as well as increased and robust recognition and categorization capabilities about the places on the map. Our future research endeavors include the integration of our system with contextual and co-occurrence constraints as well as ontological structures. Such techniques, combined with our proposed method, could endow the robots with the ability to perform more complex operations such as task planning and fetching routines, while they will be able to infer about their surroundings by collecting multiple cues.

References

[1] S. Patnaik, Robot Cognition and Navigation: An Experiment with Mobile Robots (Cognitive Technologies), Springer-Verlag, Inc., New York, 2005.
[2] S. Vasudevan, S. Gächter, V. Nguyen, R. Siegwart, Cognitive maps for mobile robots, an object based approach, Robotics and Autonomous Systems 55 (5) (2007) 359–371.
[3] P. Newman, K. Ho, Slam-loop closing with visually salient features, in: International Conference on Robotics and Automation, ICRA, IEEE, 2005, pp. 635–642.
[4] V. Chen, M. Batalin, W. Kaiser, G. Sukhatme, Towards spatial and semantic mapping in aquatic environments, in: International Conference on Robotics and Automation, ICRA, IEEE, 2008, pp. 629–636.
[5] J. Zhang, M. Marszałek, S. Lazebnik, C. Schmid, Local features and kernels for classification of texture and object categories: a comprehensive study, International Journal of Computer Vision 73 (2) (2007) 213–238.
[6] M. Ullah, A. Pronobis, B. Caputo, J. Luo, R. Jensfelt, H. Christensen, Towards robust place recognition for robot localization, in: International Conference on Robotics and Automation, ICRA, IEEE, 2008, pp. 530–537.
[7] A. Pronobis, O. Martínez Mozos, B. Caputo, P. Jensfelt, Multi-modal semantic place classification, The International Journal of Robotics Research, IJRR 29 (2–3) (2010) 298–320.
[8] A. Nuchter, K. Lingemann, J. Hertzberg, H. Surmann, 6D slam with approximate data association, in: 12th International Conference on Advanced Robotics, ICAR, IEEE, 2005, pp. 242–249.
[9] G. Grisetti, S. Grzonka, C. Stachniss, P. Pfaff, W. Burgard, Efficient estimation of accurate maximum likelihood maps in 3D, in: International Conference on Intelligent Robots and Systems, IROS, IEEE, 2007, pp. 3472–3478.
[10] K. Konolige, M. Agrawal, R. Bolles, C. Cowan, M. Fischler, B. Gerkey, Outdoor mapping and navigation using stereo vision, in: Experimental Robotics, Springer, 2008, pp. 179–190.
[11] N. Fioraio, K. Konolige, Realtime visual and point cloud slam, in: RGB-D Workshop on Advanced Reasoning with Depth Cameras at Robotics: Science and Systems Conference, RSS, vol. 27, 2011.
[12] A. Mishra, A. Shrivastava, Y. Aloimonos, Segmenting simple objects using RGB-D, in: IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2012, pp. 4406–4413.
[13] H. Jin, P. Favaro, S. Saotto, Real-time 3D motion and structure of point features: a front-end system for vision-based control and interaction, in: Conference on Computer Vision and Pattern Recognition, 2000, Vol. 2, IEEE, 2000, pp. 778–779.
[14] D. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.
[15] H. Bay, T. Tuytelaars, L. Van Gool, Surf: speeded up robust features, Computer Vision—ECCV 2006 (2006) 404–417.
[16] M. Agrawal, K. Konolige, M. Blas, Censure: center surround extremas for realtime feature detection and matching, Computer Vision—ECCV 2008 (2008) 102–115.
[17] V. Vonikakis, D. Chrysostomou, R. Kouskouridas, A. Gasteratos, Improving the robustness in feature detection by local contrast enhancement, in: International Conference on Imaging Systems and Techniques, IST, IEEE, 2012, pp. 158–163.
[18] S. Heymann, K. Muller, A. Smolic, B. Frohlich, T. Wiegand, Sift implementation and optimization for general-purpose GPU, in: International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision, 2007, pp. 317–322.
[19] R. Anati, D. Scaramuzza, K. Derpanis, K. Daniilidis, Robot localization using soft object detection, in: IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2012, pp. 4992–4999.
[20] P. Henry, M. Krainin, E. Herbst, X. Ren, D. Fox, RGB-D mapping: using depth cameras for dense 3D modeling of indoor environments, in: 12th International Symposium on Experimental Robotics, ISER, vol. 20, 2010, pp. 22–25.
[21] G. Huang, A. Mourikis, S. Roumeliotis, Analysis and improvement of the consistency of extended Kalman filter based slam, in: IEEE International Conference on Robotics and Automation, 2008, ICRA 2008, IEEE, 2008, pp. 473–479.
[22] K. Grauman, T. Darrell, The pyramid match kernel: discriminative classification with sets of image features, in: IEEE International Conference on Computer Vision, ICCV 2005, vol. 2, IEEE, 2005, pp. 1458–1465.
[23] A. Torralba, K. Murphy, W. Freeman, M. Rubin, Context-based vision system for place and object recognition, in: International Conference on Computer Vision, IEEE, 2003, pp. 273–280.
[24] C. Wallraven, B. Caputo, A. Graf, Recognition with local features: the kernel recipe, in: IEEE International Conference on Computer Vision, 2003, IEEE, 2003, pp. 257–264.
[25] J. Willamowski, D. Arregui, G. Csurka, C.R. Dance, L. Fan, Categorizing nine visual classes using local appearance descriptors, in: ICPR Workshop on Learning for Adaptable Visual Systems, 2004.
[26] D. Filliat, A visual bag of words method for interactive qualitative localization and mapping, in: International Conference on Robotics and Automation, 2007, IEEE, 2007, pp. 3921–3926.
[27] F. Fraundorfer, C. Engels, D. Nistér, Topological mapping, localization and navigation using image collections, in: International Conference on Intelligent Robots and Systems, IROS, IEEE, 2007, pp. 3872–3877.
[28] O. Linde, T. Lindeberg, Object recognition using composed receptive field histograms of higher dimensionality, in: International Conference on Pattern Recognition, ICPR, vol. 2, IEEE, 2004, pp. 1–6.
[29] E. Fazl-Ersi, J. Tsotsos, Histogram of oriented uniform patterns for robust place recognition and categorization, The International Journal of Robotics Research 31 (4) (2012) 468–483.
[30] I. Laptev, T. Lindeberg, Velocity adaptation of spatio-temporal receptive fields for direct recognition of activities: an experimental study, Image and Vision Computing 22 (2) (2004) 105–116.
[31] I. Laptev, T. Lindeberg, Local descriptors for spatio-temporal recognition, Spatial Coherence for Visual Motion Analysis (2006) 91–103.
[32] Z. Zivkovic, B. Bakker, B. Krose, Hierarchical map building using visual landmarks and geometric constraints, in: International Conference on Intelligent Robots and Systems, IROS, IEEE, 2005, pp. 2480–2485.
[33] A. Pronobis, B. Caputo, P. Jensfelt, H. Christensen, A discriminative approach to robust visual place recognition, in: International Conference on Intelligent Robots and Systems, 2006, IEEE, 2006, pp. 3829–3836.
[34] O. Mozos, C. Stachniss, W. Burgard, Supervised learning of places from range data using adaboost, in: International Conference on Robotics and Automation, ICRA, IEEE, 2005, pp. 1730–1735.
[35] A. Rottmann, O. Mozos, C. Stachniss, W. Burgard, Semantic place classification of indoor environments with mobile robots using boosting, in: National Conference on Artificial Intelligence, Vol. 20, 2005, p. 1306.
[36] R. Dirven, The Construal of Space in Language and Thought, Vol. 8, Walter de Gruyter, 1996.
[37] J. Klippenstein, H. Zhang, Quantitative evaluation of feature extractors for visual slam, in: Canadian Conference on Computer and Robot Vision, IEEE, 2007, pp. 157–164.
[38] J. Craig, Introduction to Robotics: Mechanics and Control, Prentice Hall, 2004.
[39] B. Siciliano, L. Sciavicco, L. Villani, Robotics: Modelling, Planning and Control, Springer-Verlag, 2009.
[40] R. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, Vol. 2, Cambridge Univ. Press, 2000.

[41] J. Deschaud, F. Goulette, A fast and accurate plane detection algorithm for large noisy point clouds using filtered normals and voxel growing, in: International Symposium on 3D Data Processing, Visualization and Transmission, 2010.
[42] P. Besl, N. McKay, A method for registration of 3D shapes, IEEE Transactions on Pattern Analysis and Machine Intelligence 14 (2) (1992) 239–256.
[43] S. Rusinkiewicz, M. Levoy, Efficient variants of the ICP algorithm, in: 3DIM'01, 2001, pp. 145–152.
[44] C. Yang, G. Medioni, Object modelling by registration of multiple range images, Image and Vision Computing 10 (3) (1992) 145–155.
[45] A. Jain, M. Murty, P. Flynn, Data clustering: a review, ACM Computing Surveys (CSUR) 31 (3) (1999) 264–323.
[46] T. Martinetz, S. Berkovich, K. Schulten, Neural-gas' network for vector quantization and its application to time-series prediction, Transactions on Neural Networks 4 (4) (1993) 558–569.
[47] B. Fritzke, Some competitive learning methods, in: Artificial Intelligence Institute, Dresden University of Technology.
[48] D. Nister, H. Stewenius, Scalable recognition with a vocabulary tree, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR, vol. 2, IEEE, 2006, pp. 2161–2168.
[49] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, 2000.
[50] C. Chang, C. Lin, LIBSVM: a library for support vector machines, ACM Transactions on Intelligent Systems and Technology (TIST) 2 (3) (2011) 27.
[51] Y. Liu, Y. Zheng, One-against-all multi-class svm classification using reliability measures, in: International Joint Conference on Neural Networks, IJCNN, vol. 2, IEEE, 2005, pp. 849–854.
[52] I. Kostavelis, A. Gasteratos, Cognitive navigation dataset, Group of Robotics and Cognitive Systems (2012). Available at http://robotics.pme.duth.gr/kostavelis/Dataset.html.

Ioannis Kostavelis was born in Thessaloniki, Greece, in 1987. He received the diploma degree in Production and Management Engineering from the Democritus University of Thrace and the M.Sc. degree (with Honors) in Informatics from the Aristotle University of Thessaloniki in 2009 and 2011, respectively. He is currently with the Laboratory of Robotics and Automation, Department of Production and Management Engineering, Democritus University of Thrace, where he is pursuing the Ph.D. degree in the field of robotic vision. He has been involved in different research projects funded by the European Space Agency and the Greek state. His current research interests include vision systems for robotic applications and machine learning techniques.



Antonios Gasteratos is an Associate Professor at the Department of Production and Management Engineering, Democritus University of Thrace (DUTH), Greece. He teaches the courses of Robotics, Automatic Control Systems, Measurements Technology and Electronics. He holds a B.Eng. and a Ph.D. from the Department of Electrical and Computer Engineering, DUTH, Greece. During 1999–2000 he was a visiting researcher at the Laboratory of Integrated Advanced Robotics (LIRA-Lab), DIST, University of 34 Genoa, Italy. He has served as a reviewer to numerous Scientific Journals and International Conferences. His research interests are mainly in mechatronics and in robot vision. He has published more than 140 papers in books, journals and conferences. He is a senior member of the IEEE. More details about him are available at http://robotics.pme.duth.gr/.